World J. Surg. 29, 601–605 (2005)
DOI: 10.1007/s00268-005-7921-y

WORLD
Journal of
SURGERY

Sample Size Matters: A Guide for Surgeons

Ulrich Guller, M.D., M.H.S., Daniel Oertli, M.D., F.A.C.S.

Department of Surgery, Divisions of General Surgery and Surgical Research, University of Basel, Spitalstrasse 21, CH-4031, Basel, Switzerland

Published Online: April 21, 2005

Abstract. Considerations of sample size computations in the medical literature have gained increasing importance over the past decade and are now often mandatory for scientific grant proposals, protocols, and publications. However, many surgeons are ill-prepared to understand the parameters on which the appropriate sample size is based. The present article has several objectives: first, to review the need for sample size considerations; second, to explain the ingredients necessary for sample size computations in simple, nonmathematic language; third, to provide options for reducing the sample size if it seems impracticably large; and fourth, to help avoid some of the more common mistakes encountered when computing sample sizes.

How many patients do I need to enroll in my study? This is the inevitable question that forms the preface to any clinical study. The importance of such a consideration is clear: even the most thoroughly planned and well executed investigation may fail to answer the research question if the sample size is too small. On the other hand, enrolling more patients than necessary in a study is not cost-effective and could potentially be unethical [1]. The objective of sample size computations is to estimate accurately the appropriate number of patients to be enrolled in a given study.

Performing accurate sample size computations represents a critical step in the design of scientific studies. However, some surgeons are ill-prepared to understand the various parameters that affect the sample size. The failure to carefully consider sample size and power represents a worrisome, yet frequent phenomenon [2–4]. Because simplistic and sometimes ill-informed applications of mathematic formulas are part of the problem we seek to remedy, this review provides a series of nonmathematic explanations of the key issues of sample size computations coupled with some intuitive examples drawn from the field of surgery. It is hoped that this artical facilitates surgeons' understanding of the cardinal importance of computing an appropriate sample size and as a result improve their ability to perform methodologically sound research.

Ingredients for Sample Size Computations

Here, we briefly discuss the different various parameters that affect the sample size.

Correspondence to: Ulrich Guller, M.D., M.H.S., e-mail: ugueller@uhbs.ch

One-sided versus Two-sided Hypothesis

To understand sample size computations, it is necessary to understand the meanings of the null hypothesis and the alternative hypothesis, as well as the difference between a one-sided (or one-tailed) versus two-sided (or two-tailed) hypothesis. The null hypothesis is the hypothesis that no difference exists between the study groups. In a randomized clinical trial, for example, the null hypothesis states that there is no difference between study arms for the endpoint under investigation (e.g., disease-free or overall survival, postoperative complications, postoperative mortality). Conversely, the alternative hypothesis (the one the investigator wants to prove) is that there exists a significant difference between study arms. The objective is to collect data and ascertain whether the results provide evidence against the null hypothesis.

The null and alternative hypotheses can be stated either one- or two-sided. A two-sided alternative hypothesis states that a difference exists between two treatments, with the possibility of either treatment being superior to the other. Conversely, a one-sided alternative hypothesis specifies in advance (a priori) that one treatment is believed to be superior to the other. An investigator who compares a new treatment to the standard treatment may have reason to believe that the new therapy is better based on pilot studies or data from the current literature. Choosing a one-sided alternative hypothesis increases the power for a given sample size [5]. Although two-sided alternative hypotheses are commonly used throughout the medical literature, some investigators argue that a one-sided hypothesis is more appropriate in certain situations. Here are some general guidelines regarding this issue [6].

1. Unless you can state with absolute certainty that a difference between two interventions can only go in one direction, a two-sided alternative hypothesis should be used [7]. For instance, although you might firmly believe that a new chemotherapy regimen for stage III colon cancer patients improves overall survival, patients might actually die earlier owing to unexpectedly severe side effects. Moreover, there are funding agencies that require two-sided alternative hypotheses, regardless of how strong the previous evidence is that the effect goes in only one direction.

Table 1. Type I and type II errors.

	Truth in the overall patient population			
Results of the study	**Treatment difference**		**No treatment difference**	
Treatment difference	Correct conclusion	**A**	**B**	Type I error (false-positive result)
No treatment difference	Type II error (false-negative result)	**C**	**D**	Correct conclusion

2. If you use a one-sided alternative hypothesis, it must be stated in advance (a priori hypothesis), specifying the intervention that is believed to be superior [7, 8].
3. There have been instances in medical science when, at the end of a trial yielding a marginally significant two-tailed *p*- value (based on a two-sided alternative hypothesis, e.g., $p = 0.06$) for the difference between interventions, the p-value was switched to a one-tailed *p*-value (0.03) to obtain a statistically significant result. *Such behavior is misleading and should be abandoned.*

Type I Error and Type II Error

There are two ways study findings can err. [6, 9–11].

1. The results might lead to the erroneous conclusion that a statistically significant difference exists between the study groups when in reality it does not (Table 1, cell B).
2. The results might lead to the erroneous conclusion that there is no significant difference between the study groups when in reality a difference exists (Table 1, cell C).The first situation represents false-positive result and is called a *type I error.* The bound that we put on the probability of committing a type I error is named *alpha* [9, 11]. Alpha is also referred to as the level of statistical significance or significance level. Situation 2 represents a false-negative result and is called a *type II error.* The probability of committing a type II error is referred to as *beta* [9, 11, 12].

An alpha of 0.05 is commonly assumed in medical research. This means that a 5% chance of obtaining a false-positive result (a statistically significant difference if in reality no difference exists) is considered acceptable. Alpha is the benchmark to which *p* values are compared. If the *p* value is larger than alpha, a result is said to be *nonsignificant*. On the other hand, if the *p* value is smaller than the benchmark alpha, the findings are *statistically significant.* In other words, alpha is the threshold *p* value below which a result is called statistically significant. Beta, the false-negative rate, is complementary to the power of a study. Although it could be assumed that both type I and type II errors are set at the same level of 5%, a false-positive finding is often considered potentially more harmful than a false-negative result. Thus, in medical science, beta is commonly set between 0.2 and 0.1. Type II errors of 0.2, 0.15, and 0.1 correspond to a power of 80% (1.0–0.2), 85% (1.0–0.15), or 90% (1–0.10), respectively. Ideally, both alpha and beta would be set at 0 to avoid both false-positive and false-negative findings. This would, however, result in an prohibitively large sample size, rendering any trial unfeasible. For a patient sample of a given size, there is always a trade-off between alpha and beta: the more stringent the alpha (the lower the false-positive rate), the higher the beta (increased rate of false-negative results, lower power), and vice versa [9, 13]. In general, one should choose a small alpha level if avoiding a false-positive result is particularly important (e.g., comparing the efficacy of a new chemotherapy regimen with serious adverse effects against an existing regimen that is efficient and well supported by patients). Similarly, a small beta level should be chosen if obtaining a false-negative result would be deleterious; for example, if an investigator wants to demonstrate the superiority of a new, less invasive surgical procedure over an established procedure that is associated with considerable short- and long-term sequelae. The false conclusion that the new procedure is not as effective as the standard procedure would put new patients at risk of suffering worse outcomes [6].

Power

Power is defined as the probability of finding a statistically significant result (of rejecting the null hypothesis) in a study if the populations are truly different [9]. The choice of adequate power in a study is critical, as investigators and funding agencies must be confident that an existing difference in the overall patient population can be detected using the study sample. For instance, if the power in a randomized controlled trial is set at 90% (beta of 10%) and a true difference exists between the study arms in the overall patient population, we would be able to detect that difference in 9 of 10 cases if the trial were repeated an infinite number of times. The power of a study is intrinsically linked to the sample size [14]: the larger the sample size, the higher the power. Quite often small studies do not find statistically significant differences. It is then unclear whether there is truly no difference between the treatment options or the sample size was prohibitively small to provide sufficient evidence for a statistically significant difference [4]. This represents one of the most important concepts of this article and is worth repeating: To interpret appropriately an investigation that does not find a statistically significant difference between the study groups, it is critically important to check if a sample size was computed and whether a sufficient amount of patients were enrolled and evaluated [15]. Raising the following question is imperative: Based on the actual sample size, the observed difference between the study groups, and the chosen alpha level, what was the power of the investigation [1, 5]? Studies that neither found statistically significant difference between the outcomes nor enrolled a sufficiently large number of patients should be considered *inconclusive* (not *negative*), and the results should be interpreted (at best) as hypothesis generating but not as hypothesis testing [15]. Unfortunately, there is a plethora of studies in the medical literature that were clearly underpowered while claiming that there was no statistically significant difference in outcomes [2, 3, 12, 15, 16], an erroneous and potentially harmful conclusion. Goodman and Berlin emphasized the importance of displaying 95% confidence intervals of the observed difference between groups to enable the reader to more accurately interpret "negative" studies [17].

Table 2. Sample size computations for a hypothetical randomized clinical trial comparing surgery alone versus neoadjuvant radiotherapy plus surgery for esophageal cancer.

Expected 2-year overall survival of patients undergoing surgery alone (%)	Expected 2-year overall survival of patients undergoing neoadjuvant radiotherapy plus surgery (%)	Alpha	Power (%)	Total sample size
40	80	0.05	80	32
40	80	0.05	90	44
40	70	0.05	80	63
40	70	0.05	90	84
40	60	0.05	80	146
40	60	0.05	90	196
40	50	0.05	80	587
40	50	0.05	90	785
40	45	0.05	80	2324
40	45	0.05	90	3111

All sample size computations are based on equal percentages of patients in both groups, a type I error probability of 0.05 (two-sided), a projected accrual period of 2 years, and a minimum follow-up period of 2 years.

There is suggestive evidence in the medical literature that neoadjuvant radiotherapy for esophaged cancer patients might be beneficial. Let us say that we want to design a study evaluating whether adjuvant radiotherapy plus surgery prolongs overall survival versus surgery alone. Table 1 displays sample size computations that would answer the same research question using different overall survival and power estimates. It is important to realize that the sample sizes are highly dependent on the assumed estimated difference in survival rate and the chosen power.

In addition to the sample size, the power of a study depends on the following factor [9, 12].

1. The extent of the true difference in the overall patient populations from which the sample has been drawn
2. The alpha level (accepted rate of false-positive results)
3. Variability (standard deviation) of continuous outcomesThe power of the study increases with larger sample sizes, larger true differences in the outcomes between the populations from which the patient samples have been drawn, higher acceptance of false-positive results, and less variability in the outcomes.

Effect Size

For sample size computations, investigators start by defining a clinically meaningful difference in outcomes (synonym: endpoints) between treatment A and B (called effect size or *delta*), which is believed to be true for the overall patient population [14, 18]. Although a study often evaluates different outcomes, one should be chosen for consideration as the most relevant one (the primary outcome, e.g., overall survival in a randomized phase III trial); and the sample size computations must be based on an estimated difference of the primary endpoint. On rare occasions, two or even more outcomes are considered to be of similar relevance. In that scenario, the sample size should be computed for all outcomes and the largest resulting sample size chosen for the study [5, 15].

It is often a challenging undertaking to estimate the magnitude of the difference in outcomes because the true value of the effect size might be totally unknown [19]. Clearly, if the effect size were already known, there would be no point in performing the study. Ideally, the estimated difference between outcomes is based on the preliminary data gained from pilot studies [20] or retrospective reviews, but is sometimes specified according to clinical intuition [1]. The smaller the expected difference in the outcome, the larger is the sample size (Table 2). However, this difference should be sufficiently large to result in a change in clinical practice [18]. For instance, suppose a study evaluating whether pancreaticoduodenectomy plus extended lymphadenectomy (the new treatment) leads to improved overall survival compared to pancreaticoduodenectomy alone (standard treatment) in patients with potentially curable pancreatic cancer. Let us further assume that the median overall survival after the standard treatment is known to be 24 months. It would certainly be irrelevant if the new, more extensive surgical therapy led to a median overall survival of 25 months. Suppose, however, that the investigator believes that the median overall survival due to more extensive surgery will be 36 months. This would represent a clinically relevant difference that would affect the current standard of care. The investigator would then power the study for a difference of 12 months (36 − 24 months).

Variability in Outcomes

If the endpoint under investigation is a continuous variable, the investigator must define not only a clinically relevant difference between the outcomes but also the variability (standard deviation, data scatter) of the outcome. The larger the variability in outcomes, the more patients needed to prove that the difference in outcomes is statistically significant (Table 3). As for the estimation of difference in outcomes, the assumption of outcomes variability ideally is based on data of previous studies of similar patient populations or data from the medical literature. This obviously does not apply to comparisons of percentages.

How to Minimize Sample Sizes?

It is intuitively obvious that large sample sizes are unequivocally associated with increased costs and consumption of resources when performing clinical trials. What can you do if after computing the sample size based on the above-mentioned factors, you realize that it is impracticably large? To answer this relevant question it is worth reviewing the parameters affecting the sample size. The sample size of a study is large if [5]

1. The estimated effect size is small
2. The type I error (rate of false positives) is small
3. The type II error (rate of false negatives) is small (equivalent to high power)
4. The estimated variability of the outcomes is large
6. The alternative hypothesis is formulated two-sidedKnowing the association of these parameters and the sample size, the

Table 3. Hypothetical randomized clinical trial comparing quality of life 2 weeks postoperatively in patients randomized to open versus laparoscopic sigma resection for diverticulitis.

QoL open	QoL lap	SD	Power (%)	Hypothesis	Total no. of patients
50	60	5	80	Two-sided	12
50	60	5	80	One-sided	10
50	60	15	80	Two-sided	74
50	60	15	80	One-sided	58
50	60	30	80	Two-sided	286
50	60	30	80	One-sided	226

Suppose that we want to design a prospective randomized trial comparing quality of life 2 weeks postoperatively in patients randomized to open versus laparoscopic sigma resection for diverticulitis. The quality of life assessments are done on an imaginary questionnaire ranging from 0 (worst imaginable quality of life) to 100 (excellent quality of life).

Note that the sample size increases with greater data scatter (standard deviation) and decreases if the hypothesis is one-sided.

QoL: quality of life; open: open operation; Lap: laparoscopic surgery.

investigator should critically review whether one or several of them could be changed. Is the estimated effect size unreasonably small? Could the type I or type II error be increased without either putting patients at risk or affecting the methodologic rigor of the study? Can the measurements be done with greater precision to diminish the variability (noise) in outcomes? Is there sound evidence that allows the choice of a one-sided instead of a two-sided alternative hypothesis?

If none of the above-mentioned parameters can be changed, the following alternatives was result in a decreased sample size.

1. *Use continuous instead of dichotomous* (*yes/no*) *endpoints:* Some variables can be expressed as either continuous or dichotomous (yes/no) endpoints. If the choice of a continuous endpoint is possible, it usually results in a smaller sample size compared with dichotornous endpoints [11], as the collapse of continuous data into two categories is associated with loss of information. Suppose that in a randomized clinical trial we are comparing two neoadjuvant treatments for advanced-stage rectal cancer and that one outcome is the carcinoembryonic antigen (CEA) level after therapy. If we compare the percentage of patients with normal CEA levels (a dichotomous outcome: normal versus abnormal), the sample size is larger than if absolute CEA levels (continuous outcome) are compared.
2. *Use a paired instead of an unpaired data analysis:* A paired (or clustered) data analysis should be used in the following situations: 2 (a) if two or more measurements are done with the same subjects (e.g., before and after an intervention) [20]; (b) if subjects are matched pairs or clusters (e.g., if for each patient in a certain sample another patient with similar characteristics (e.g., age, gender, race) has been assigned in a comparison sample, or if there are natural clusters such as members within families) [20]. The use of a paired data analysis results in smaller sample sizes compared with nonpaired data analysis. The mathematical explanation of this phenomenon is beyond the scope of the present article. However, it is intuitive that if each patient serves as his/her own control or that if you compare patients with similar characteristics (in the matched study design), the between-patient variability is decreased. In paired data situations a paired test (e.g., paired *t*-test, repeated-measures ANOVA, McNemar's test.) may be used yielding greater power than with the corresponding nonpaired test.
3. *Use a more common outcome or a compared outcome:* The power of the study depends more on the number of patients with a certain outcome (event) than on the total number of subjects [15, 21]. When the computed sample size seems prohibitively large, one could attempt to choose a more common outcome. A simple way to increase the occurrence of an outcome (e.g., death) is to enroll patients with a higher risk of experiencing this outcome. If a researcher evaluates a new investigational therapy for melanoma, and overall survival is the primary endpoint, she could choose to enroll only stage III and IV patients who are at greater risk of experiencing the outcome (death) than are early-stage patients. Moreover, the number of events could be increased by extending the period of follow-up, which again lowers the sample size [22, 23]. However, prolonging the follow-up of patients leads to higher costs and may be associated with an increased rate of dropouts. Finally, a composite outcome could be chosen. Suppose that the primary endpoint of a trial comparing the impact of two treatments in patients with metastatic colon cancer is disease regression. Further suppose that the sample size was prohibitively large to perform the trial. To increase the feasibility of the study, one could choose a composite outcome, such as (disease regression + stable disease).It should be emphasized, however, that an investigator should only change the parameters affecting the sample size if there is evidence to support this change. Decreasing the sample size for the sake of feasibility at the expense of making unrealistic assumptions or compromising the relevance of the research question results in an underpowered or irrelevant trial that is both wasteful and unethical.

Caveats of Sample Size Computations

Performing sample size computations represents an important and delicate step when designing a study. Herein, we briefly summarize some caveats regarding sample size computations.

1. The most common error regarding sample size computations is performing them too late in the process of developing the study. Sample size computations should be done during the very early planning phases of a study, when fundamental changes of the design are still possible [5]. Also, sample size computations often reveal that the number of patients needed is impracticably large, thus compromising the feasibility of the investigation.
2. It might be wise to include more patients than the minimum number of participants computed. This depends on the estimated percentage of dropouts or those lost-to-follow-up. The

sample size computations reflect the number of *evaluable* patients at the end of the study. If the sample size computations do not take this into consideration, the study is underpowered if some patients do not finish the trial.

3. It is imperative that the authors of a clinical trial report the parameters upon which the computed sample size is based [15, 16, 24]. Despite this, many investigators fail to do so [2, 3, 12, 25]. If no information about power calculations is reported, the reader does not know if: (a) no sample size requirement was computed; (b) the investigators were unable to accrue the initially computed patient number; (c) the trial was extended beyond the initially computed sample size to obtain higher statistical power; or (d) the investigators stopped the trial earlier than anticipated because the interim results were favorable [25].
4. An important caveat regarding sample size computations represents the misinterpretation between number of patients and number of events [15]. Some mathematic formulas provide the investigator with the number of events (e.g., number of needed death in a trial evaluating overall survival) rather than with the total number of patients needed. Let us assume that we estimate that the overall survival rate is 30% at the end of a planned trial and that the number of events needed to show a statistically significant difference is 200. If we misinterpreted this number as the total number of patients, we would end up with a dramatically underpowered study, a grave error that must be avoided at any costs. Verifying sample size calculations with a statistician is strongly recommended to double check whether the sample size refers to total number of patients or number of required events.

Conclusions

Understanding the parameters that affect sample sizes is invaluable in the design of clinical studies as well as essential to critical assessment of scientific findings and their implementation in clinical practice. It is hoped that this review enables surgeons to understand the ingredients necessary for accurate sample size computations and facilitates their communication with medical epidemiologists and statisticians, with the ultimate goal of creating better surgical trials. Clearly, sample size matters.

Acknowledgments

The authors thank Jonathan McCall for carefully reading the manuscript and making many valuable suggestions.

References

1. Lerman J. Study design in clinical research: sample size estimation and power analysis. Can. J. Anaesth 1996;43:184–191
2. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122–124
3. Vandekerckhove P, O'Donovan PA, Lilford RJ, et al. Infertility treatment: from cookery to science: The epidemiology of randomised controlled trials. Br. J. Obstet. Gynaecol 1993;100:1005–1036
4. Williams HC, Seed P. Inadequate size of negative' clinical trials in dermatology. Br. J. Dermatol 1993;128:317–326
5. Sheps S. Sample size and power. J. Invest. Surg 1993;6:469–475
6. Guller, U, Delong, ER (2004) "Interpreting statistics in medical Literature: A vade mecum for surgeons." J Am Coll Surgeons 198: 441
7. Bland JM, Altman DG. One and two sided tests of significance. B.M.J. 1994;.:309–248
8. O'Brien PC, Shampo MA. Statistics for clinicians. 8. Comparing two proportions: the relative deviate test and chi-square equivalent. Mayo. Clin. Proc. 1981;56:513–515
9. Berwick DM. Experimental power: the other side of the coin. Pediatrics 1980;65:1043–1045
10. Goodman SN. Toward evidence-based medical statistics. 1. The P value fallacy. Ann. Intern. Med 1999;130:995–1004
11. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians. 1. Hypothesis testing. Csan. Med. Assac. J. Grasi 1995;152:27–32
12. Freiman JA, Chalmers TC, Smith H Jr, et al. The importance of beta, the type II error and sample size in the desigrand interpretation of the randomized control trial: survey of 71 ''negative'' trials. N. Engl. J. Med 1978;299:690–694
13. Perneger TV. What's wrong with Bonferroni adjustments. B. M. J. 1998;316:1236–1238
14. Pascoe JM. Was it a type II error?. Pediatrics 1981;68:149–150
15. Fayers PM, Machin D. Sample size: how many patients are necessary? . Br J. Cancer 1995;72:1–9
16. Altman DG. Statistical reviewing for medical journals. Stat. Med. 1998;17:2661–2674
17. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann. Intem. Med. 1994;121:200–206
18. Lilford R, Braunholtz D, Harris J, et al. Trials in surgery. Br. J. Surg. 2004;91:6–16
19. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Stat. Med. 1990;9:65–72
20. O'Brien PC, Shampo MA. Statistics for clinicians. 5. One sample of paired observations (paired t test). Mayo Clin-Proc. 1981;56:324–326
21. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br. J. Cancer 1976;34:585–612
22. Church TR, Ederer F.andel JS, et al. Estimating the duration of ongoing prevention trials. Am. J. Epidemiol 1993;137:797–810
23. Ederer F, Church TR, Mandel JS. Sample sizes for prevention trials have been too small. Am. J. Epidemiol 1993;137:787–796
24. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br. J. Cancer 1994;69:979–985
25. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. N. Engl. J. Med 1987;317:426–432